Complete genome sequence of *Saccharomonospora viridis* type strain (P101^T)

Amrita Pati[1], Johannes Sikorski[2], Matt Nolan[1], Alla Lapidus[1], Alex Copeland[1], Tijana Glavina Del Rio[1], Susan Lucas[1], Feng Chen[1], Hope Tice[1], Sam Pitluck[1], Jan-Fang Cheng[1], Olga Chertkov[1,3], Thomas Brettin[1,3], Cliff Han[1,3], John C. Detter[1,3], Cheryl Kuske[1,3], David Bruce[1,3], Lynne Goodwin[1,3], Patrick Chain[1,4], Patrik D'haeseleer[1,4], Amy Chen[5], Krishna Palaniappan[5], Natalia Ivanova[1], Konstantinos Mavromatis[1], Natalia Mikhailova[1], Manfred Rohde[6], Brian J Tindall[2], Markus Göker[2], Jim Bristow[1], Jonathan A. Eisen[1,7], Victor Markowitz[5], Philip Hugenholtz[1], Nikos C. Kyrpides[1], and Hans-Peter Klenk[2*]

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[3] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico USA
[4] Lawrence Livermore National Laboratory, Livermore, California 94550, USA
[5] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[6] HZI - Helmholtz Centre for Infection Research, Braunschweig, Germany
[7] University of California Davis Genome Center, Davis, California, USA

[*]Corresponding author: Hans-Peter Klenk

Keywords
thermophile, hot compost, Gram-negative actinomycete, farmer's lung disease, bagassosis, humidifier fever, pentachlorophenol metabolism, *Pseudonocardiaceae*

Abstract
Saccharomonospora viridis (Schuurmans *et al.* 1956) Nonomurea and Ohara 1971 is the type species of the genus *Saccharomonospora* which belongs to the family *Pseudonocardiaceae*. *S. viridis* is of interest because it is a Gram-negative organism classified amongst the usually Gram-positive actinomycetes. Members of the species are frequently found in hot compost and hay, and its spores can cause farmer's lung disease, bagassosis, and humidifier fever. Strains of the species *S. viridis* have been found to metabolize the xenobiotic pentachlorophenol (PCP). The strain described in this study has been isolated from peat-bog in Ireland. Here we describe the features of this organism, together with the complete genome sequence, and annotation. This is the first complete genome sequence of the family *Pseudonocardiaceae*, and the 4,308,349 bp long single replicon genome with its 3906 protein-coding and 64 RNA genes is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project.

Introduction
Saccharomonospora viridis strain P101^T (DSM 43017 = ATCC 15386 = JCM 3036 = NCIMB 9602) is the type strain of the species, which represents the type species of the genus *Saccharomonospora* [1, 2], which presently contains nine species [3]. Although phylogenetically a member of the Gram-positive actinomycetes, already the initial report on *S. viridis* strain P101^T noticed the astonishing feature of the organism to be Gram-negative,

despite showing the typical mycelium morphology of *Saccharomonospora* [2]. Like in other actinomycetes, spores of *S. viridis* are readily dispersed in air, and apparently the prolonged exposure to the antigens of spores can result in acute respiratory distress (farmer's lung disease) which may lead to irreversible lung damage [4, 5].

Here we present a summary classification and a set of features for *S. viridis* P101^T (Table 1), together with the description of the complete genomic sequencing and annotation.

Classification and features of organism

Members of the species *S. viridis* have been isolated or molecularly identified on several occasions from hot composts in Europe and USA, [6-9], and also from soil in Japan [1]. One novel, yet unpublished, cultivated member of the species has been reported by Lu and Liu from Chinese soil (AF127525). Uncultured clone sequences with significant (99%) sequence similarity were observed from composting mass in China (AM930281 and AM930338). Screening of environmental genomic samples and surveys reported at the NCBI BLAST server indicated no closely related phylotypes that can be linked to the species or genus, with the closest matches (about 90% sequence similarity) to strain P101^T 16S rRNA identified in a marine metagenome from the Sargasso Sea [10].

Figure 1 shows the phylogenetic neighborhood of *S. viridis* strain P101^T in a 16S rRNA based tree. The sequences of all three copies of the 16S rRNA gene are identical and perfectly match the previously published 16S rRNA sequence generated from NCIMB 9602 (Z38007).

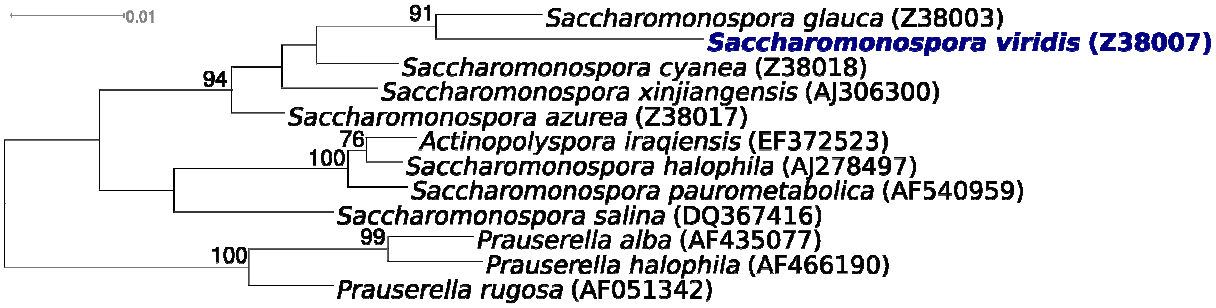
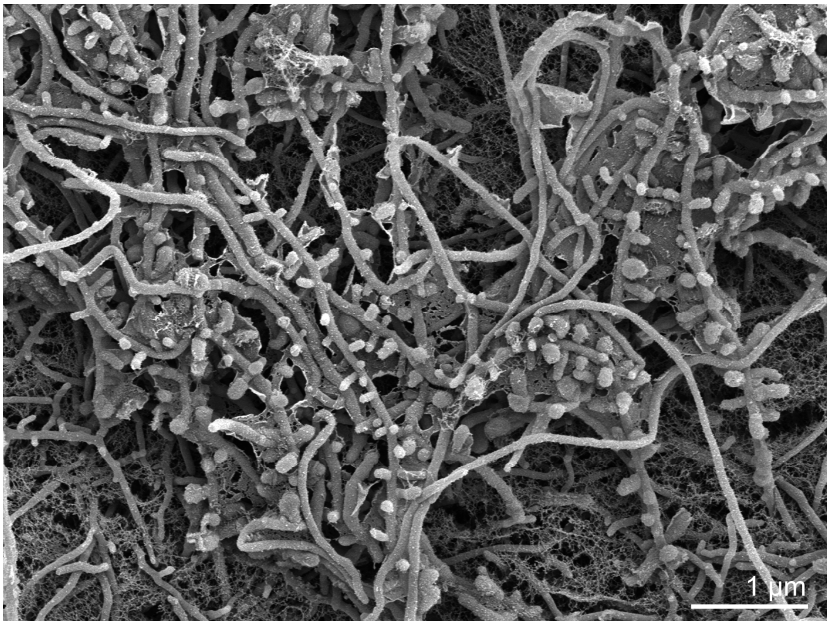


Figure 1. Phylogenetic tree of *S. viridis* strain P101^T and all type strains of the genus *Saccharomonospora* inferred from 1474 aligned characters [11, 12] of the 16S rRNA gene under the maximum likelihood criterion [13]. The tree was rooted with all type strains of the members of the genus *Prauserella*, another genus in the family *Pseudonocardiaceae*. The branches are scaled in terms of the expected number of substitutions per site. Numbers above branches are support values from 1000 bootstrap replicates if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [14] are shown in blue, published genomes in bold.

The hyphae of the vegetative mycelium of strain P101^T are branched and sometimes show curved endings [6]. Single spores are observed only on the aerial mycelium either directly on the hyphae or on short sporophores (Figure 2). The spores are oval, 0.9-1.1µm × 1.2-1.4 µm in size. Only very occasionally two spores are observed. The aerial mycelium is either greyish green in colour, or turns from white to greenish as on Czapek Agar. The optimal temperature for growth is 55°C, but 45°C for aerial mycelium formation and pigment production. At 37°C and 60°C the growth is very limited and without aerial mycelium. No growth occurs at 27°C and 70°C [6].

Strain P101^T has been observed to be sensitive to a variety of phages [15]. Members of *S. viridis* are apparently able to metabolize pentachlorophenol but not other chlorophenols [9]. It was suggested that *S. viridis* metabolizes PCP by conjugation to form a more polar transformation product, but unlike other PCP-degrading bacteria, is incapable of effecting total degradation of the xenobiotic [9]. Microorganisms such as *S. viridis* may therefore contribute to PCP removal by microbial communities in situ, despite being unable to completely mineralize chlorophenols in pure culture [9]. *S. viridis* produces a thermostable α-amylase which forms 63% (w/w) maltose on hydrolysis of starch [16]. Maltotriose and maltotetraose are the only intermediate products observed during this reaction, with maltotriose accumulating to 40% (w/w). Both unimolecular and multimolecular mechanisms (transfers and condensation) have been shown to occur during the concentration-dependent degradation of maltotriose and maltotetraose. Such reactions result in the almost exclusive formation of maltose from maltotriose at high initial concentration [16]. *S. viridis* produces thermoviridin, an antibiotic that is primarily active against the Gram-positive bacteria (growth inhibition) [15, 2]. At higher concentrations, also Gram-negative bacteria were growth-inhibited [2].

Figure 2. Scanning electron micrograph of *S. viridis* P101^T



***Chemotaxonomy*.** The murein of P101^T is of cell wall type IV. It contains meso-diaminopimelic acid in the peptidoglycan and arabinose and galactose in whole-cell hydrolysates (sugar type A). Mycolic acids and teichonic acids were not reported. Strain P101^T contains menaquinones MK-9(H_4) (60%) and MK-8(H_4) (20 to 30%). The combination of the tetrahydromultiprenyl menaquinones MK-9(H_4) and MK-8(H_4) is characteristic for the genus *Saccharomonospora* [15]. The major cellular fatty acids are saturated, iso-branched acids with 16 and 18 carbon atoms, and 2-hydroxydodecanoic acids. Details are described in the Compendium of *Actinobacteria* [17]. Phosphatidylethanolamine, hydroxy-phosphatidyl-ethanolamine, and lyso-phosphatidyl-ethanolamine were identified as the main phospholipids.

Table 1. Classification and general features of *S. viridis* P101^T in accordance with the MIGS recommendations [18]

MIGS ID	Property	Term	Evidence code[a,b]
	Current classification	Domain *Bacteria*	
		Phylum *Actinobacteria*	
		Class *Actinobacteria*	TAS [19]
		Order *Actinomycetales*	TAS [19]
		Suborder *Pseudonocardineae*	TAS [19]
		Family *Pseudonocardiaceae*	TAS [19]
		Genus *Saccharomonospora*	TAS [1]
		Species *Saccharomonospora viridis*	TAS [2]
		Type strain P101	
	Gram stain	negative	TAS [2]
	Cell shape	variable	TAS [17]
	Motility	nonmotile	NAS
	Sporulation	single spores on mainly aerial mycelium	TAS [1]
	Temperature range	thermophile, 37-60°C	TAS [15]
	Optimum temperature	55°C for growth, 45°C for aerial mycelium formation	TAS [1, 6, 15]
	Salinity	7% NaCl	TAS [15]
MIGS-22	Oxygen requirement	aerobic; nor reported if essential	TAS [15]
	Carbon source	D-glucose, sucrose, dextrin	TAS [15]
	Energy source	carbobydrates	TAS [15]
MIGS-6	Habitat	peat and compost (species occurrence)	TAS [1, 4, 6, 8, 9]
MIGS-15	Biotic relationship	Free living	
MIGS-14	Pathogenicity	Lung damage	TAS [4]
	Biosafety level	1	TAS [20]
	Isolation	peat-bog at 250 cm depth	TAS [6]
MIGS-4	Geographic location	Irish peat	
MIGS-5	Sample collection time	before 1963	TAS [6]
MIGS-4.1 MIGS-4.2	Latitude – Longitude	not reported	
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

a) Evidence codes - IDA: Inferred from Direct Assay (first time in publication); TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from http://www.geneontology.org/GO.evidence.shtml of the Gene Ontology project [21]. If the evidence code is IDA, then the property should have been directly observed, for the purpose of this specific publication, for a live isolate by one of the authors, or an expert or reputable institution mentioned in the acknowledgements.

Genome sequencing and annotation information

Genome project history

This organism was selected for sequencing on the basis of each phylogenetic position, and is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project. The genome project is deposited in the Genome OnLine Database [14] and the complete genome sequence in GenBank NOT YET. Sequencing, finishing and annotation were performed by the DOE Joint Genome Institute (JGI). A summary of the project information is shown in Table 2.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished

MIGS-28	Libraries used	Two Sanger libraries - 8 kb pMCL200 and fosmid pcc1Fos
MIGS-29	Sequencing platforms	ABI3730
MIGS-31.2	Sequencing coverage	12.9x Sanger
MIGS-30	Assemblers	phrap
MIGS-32	Gene calling method	Genemark 4.6b, tRNAScan-SE-1.23, infernal 0.81
	INSDC / Genbank ID	not yet available
	Genbank Date of Release	not yet available
	GOLD ID	Gi02228
	NCBI project ID	20835
	Database: IMG-GEBA	2500901760
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation

S. viridis strain P101^{T}, DSM 43017, was grown in DSMZ medium 83 (CZAPEC PEPTONE Medium) at 45°C. DNA was isolated from 1-1.5 g of cell paste using Qiagen Genomic 500 DNA Kit (Qiagen, Hilden, Germany) with a modified protocol for cell lysis including overnight incubation with lysozyme, mutanolysine, lysostaphine, achromopeptidase, and proteinase K at 35°C.

Genome sequencing and assembly

The genome was sequenced using Sanger sequencing platform only. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov. The Phred/Phrap/Consed software package (www.phrap.com) was used for sequence assembly and quality assessment. After the shotgun stage reads were assembled with parallel phrap (High Performance Soft ware, LLC). Possible mis-assemblies were corrected with Dupfinisher [22] or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, custom primer walk or PCR amplification (Roche Applied Science, Indianapolis, IN). A total of 354 finishing reactions were produced to close gaps and to raise the quality of the finished sequence. The completed genome sequences of *S. viridis* contains 66,210 Sanger reads, achieving an average of 12.9x sequence coverage per base, with an error rate less than 1 in 100,000.

Genome annotation

Genes were identified using GeneMark [23] as part of the genome annotation pipeline in the Integrated Microbial Genomes Expert Review (IMG-ER) system (http://img.jgi.doe.ogv/er) [24], followed by a round of manual curation using JGI's GenePRIMP pipeline (http://geneprimp.jgi-psf.org) [25]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. The tRNAScanSE tool [26] was used to find tRNA genes, whereas ribosomal RNAs were found by using the tool RNAmmer [26]. Other non coding RNAs were identified by searching the genome for the Rfam profiles using INFERNAL (v0.81) [28]. Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform (http://img.jgi.doe.gov/) [29].

Metabolic network analysis

The metabolic Pathway/Genome Database (PGDB) was computationally generated using Pathway Tools software version 12.5 [30] and MetaCyc version 12.5 [31], based on annotated EC numbers and a customized enzyme name mapping file. It has undergone no subsequent manual curation and may contain errors, similar to a Tier 3 BioCyc PGDB [32].

Genome properties

The genome is 4,308,349 bp long and comprises one main circular chromosome with a 67.3% GC content (Table 3). Of the 3970 genes predicted, 3906 were protein coding genes, and 64 RNAs. 78 pseudogenes were also identified. 71.2% of the genes were assigned with a putative function while the remaining are annotated as hypothetical proteins. The properties and the statistics of the genome are summarized in Table 3. The distribution of genes into GOGs functional categories is presented in Table 4.

Table 4. Genome Statistics

Attribute	Value	% of Total
Genome size (bp)	4,308,349	
DNA Coding region (bp)	3,805,483	88.33%
DNA G+C content (bp)	2,900,171	67.32%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	3970	
RNA genes	64	1.61%
rRNA operons	3	
Protein-coding genes	3906	98.39%
Pseudo genes	78	1.96%
Genes with function prediction	2828	71.23%
Genes in paralog clusters	534	13.45%
Genes assigned to COGs	2709	68.24%
Genes assigned Pfam domains	2845	71.66%
Genes with signal peptides	725	18.26%
Genes with transmembrane helices	880	22.17%
CRISPR repeats	9	

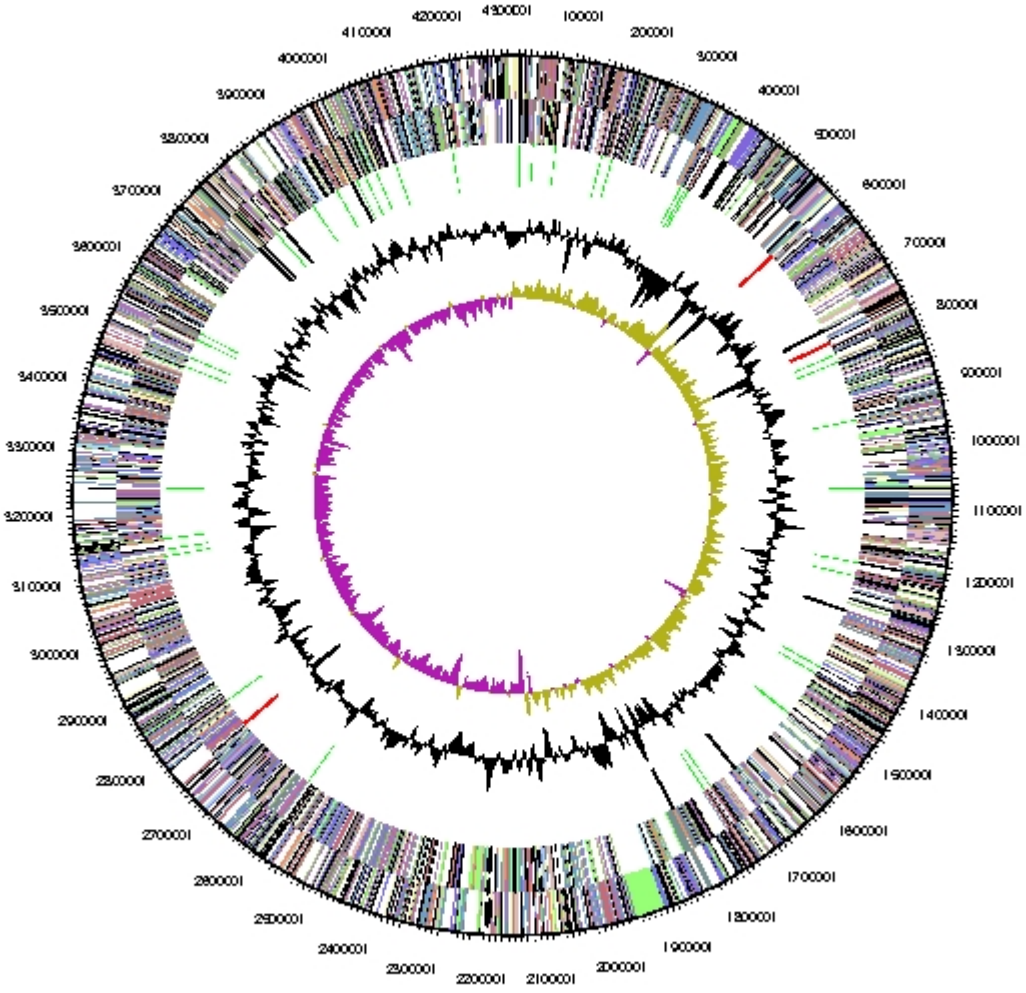


Figure 3. Graphical circular map of the genome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, sRNAs red, other RNAs black), GC content, GC skew.

Table 4. Number of genes associated with the 21 general COG functional categories

Code	COG counts and percentage of protein-coding genes		Description
	Genome		
	value	% of total	
J	158	4.0	Translation, ribosomal structure and biogenesis
A	1	0.0	RNA processing and modification
K	276	7.1	Transcription
L	125	3.2	Replication, recombination and repair
B	1	0.0	Chromatin structure and dynamics
D	25	0.6	Cell cycle control, mitosis and meiosis
Y	0	0.0	Nuclear structure

V	44	1.1	Defense mechanisms
T	146	3.7	Signal transduction mechanisms
M	125	3.2	Cell wall/membrane biogenesis
N	2	0.1	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	27	0.7	Intracellular trafficking and secretion
O	107	2.7	Posttranslational modification, protein turnover, chaperones
C	214	5.5	Energy production and conversion
G	214	5.5	Carbohydrate transport and metabolism
E	293	7.5	Amino acid transport and metabolism
F	85	2.2	Nucleotide transport and metabolism
H	175	4.5	Coenzyme transport and metabolism
I	189	4.8	Lipid transport and metabolism
P	146	3.7	Inorganic ion transport and metabolism
Q	139	3.6	Secondary metabolites biosynthesis, transport and catabolism
R	389	10.0	General function prediction only
S	182	4.7	Function unknown
-	1197	30.6	Not in COGs

Figure 4. Cellular overview diagram**.** This diagram provides a schematic of all pathways of *Saccharomonospora viridis* strain P101 metabolism. Nodes represent metabolites, with shape indicating class of metabolite (see key to right). Lines represent reactions.

Table 5. Metabolic Network Statistics

Attribute	Value
Total genes	3970
Enzymes	880
Enzymatic reactions	1155
Metabolic pathways	244
Metabolites	863

Acknowledgements

We would like to gratefully acknowledge the help of Marlen Jando for growing *S. viridis* cultures and Susanne Schneider for DNA extraction and quality analysis (both at DSMZ). This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396, as well as German Research Foundation (DFG) INST 599/1-1.

References

1. Nonomura H, Ohara Y. Distribution of Actinomycetes in soil. (X) New genus and species of monosporic actinomycetes. *J Ferment Technol* 1971, **49:**895-903.

2. Schuurmans DM, Olson BH, San Clemente CL. Production and isolation of thermoviridin, an antibiotic produced by *Thermoactinomyces viridis* n. sp. *Appl Environ Microbiol* 1956, **4:**61-6.

3. Euzéby JP. List of bacterial names with standing in nomenclature: a folder available on the internet. *Int J Syst Bacteriol* 1997, **47:**590-2.

4. Amner W, Edwards C, McCarthy AJ. Improved medium for recovery and enumeration of the farmer's lung organism, *Saccharomonospora viridis*. *Appl Environ Microbiol* 1989, **55:**2669-74.

5. Roussel S, Reboux G, Dalphin J-C, Pernet D, Laplante J-J, Millon L, Piarroux R. Farmer's Lung Disease and Microbiological Composition of Hay: A Case Control Study. *Mycopathologia* 2005, **160:**273-9.

6. Küster E, Locci R. Studies on peat and peat microorganisms. I. Taxonomic studies on thermophilic *Actinomycetes* isolated from peat. *Arch Microbiol* 1963, **45:**188-97.

7. Dees PM. WCG. Microbial diversity in hot synthetic compost as revealed by PCR-amplified rRNA sequences from cultivated isolates and extracted DNA. *FEMS Microbiol Ecol* 2001, **35:**207-16.

8. Steger K, Jarvis Å, Vasara T, Romantschuk M, Sundh I. Effects of differing temperature management on development of Actinobacteria populations during composting. *Res Microbiol* 2007, **158:**617-24.

9. Webb MD, Ewbank G, Perkins J, McCarthy AJ. Metabolism of pentachlorophenol by *Saccharomonospora viridis* strains isolated from mushroom compost. *Soil Biol Biochem* 2001, **33:**1903-14.

10. Venter JC, Remington K, Heidelberg J, Halpern A, Rusch D *et al* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004, **304:**66-74.

11. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002, **18**:452-64.

12. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000, **17**:540-52.

13. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web-servers. *Syst Biol* 2008, **57**:758-71.

14. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes OnLine Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata *Nucleic Acids Res* 2008, **36**:D475-9.

15. Greiner-Mai E, Korn-Wendisch F, Kutzner HJ. Taxonomic revision of the genus *Saccharomonospora* and description of *Saccharomonospora glauca* sp. nov. *Int J Syst Bacteriol* 1988, **38:**398-405.

16. Fogarty WM, Collins BS, Doyle EM, Kelly CT. The high maltose-forming a-amylase of *Saccharomonospora viridis*: mechanisms of action. *J Ind Microbiol Biotechnol* 1993, **11:**199-204.

17. Wink JM. Compendium of *Actinobacteria*. http://www.gbif-prokarya.de/micro-organisms/wink_pdf/DSM43017.pdf 2009.

18. Field D, Garrity G, Gray T, Morrison N, Selengut J, *et al* Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. *Nat Biotechnol* 2008, **26**:541-7.

19. Stackebrandt E, Rainey FA, Ward-Rainey NL. Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. *Int J Syst Bacteriol* 1997, **47:**479-91.

20. Biological Agents: Technical rules for biological agents www.baua.de TRBA 466.

21. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet* 2000, **25**:25-9.

22. Sims D, Brettin T, Detter JC, Han C, Lapidus A *et al*. Complete genome of *Kytococcus sedentarius* type strain (strain 541^{T}). *SIGS*, 2009 reviewed.

23. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 2001, **29**:2607-18.

24. Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu K *et al*. Expert Review of Functional Annotations for Microbial Genomes. *Submitted* 2009.

25. Pati *et al*. GenePRIMP: A Gene Prediction Improvement Pipeline for microbial genomes. *in preparation* 2009

26. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997, **25**:955-64.

27. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007, **35**: 3100-8.

28. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 2005, **33**:D121-4.

29. Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K *et al*. The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res* 2008, **36:**D528-33.

30. Karp PD, Paley SM, Romero P. The Pathway Tools Software. *Bioinformatics* 2000, **18**:S225-32.

31. Karp P, Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley SM, Rhee SY, Shearer A, Tissier C, Walk TC, Zhang P. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/Genome Databases. *Nucleic Acids Res* 2008, **36**:D623-31.

32. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005, **33**:6083-6089.